ANOTHER LOOK AT THE RETINA AS AN IMAGE SCALAR QUANTIZER

Khaled Masmoudi, Marc Antonini, Pierre Kornprobst

Another look at retina as an image scalar quantizer

Khaled Masmoudi, Marc Antonini
Univ. Nice Sophia Antipolis-CNRS-I3S
Sophia Antipolis, France
kmasmoud, am@i3s.unice.fr

Pierre Kornprobst
INRIA-NeuroMathComp
Sophia Antipolis, France
pierre.kornprobst@sophia.inria.fr

***Abstract*— We present a novel quantization algorithm based on a biologically realistic vertebrate retina model. Our aim is to investigate the retina transfer function that maps a continuous input signal into a binary-like neural code. We, furthermore, describe a possible decoding procedure. The coder/decoder, we describe here, focuses on the temporal behavior of the three last retina layers. The neural code, that is originated by these layers, contains all the information required about the stimulus. Actually, the considered code is a series of electrical impulses, termed as *spike trains* [1]. The coding/decoding schema we introduce assumes that, given a ΔT-sized time bin, the count of spikes convey the major part of the information. This defines a so-called *firing rate coding* which is the most commonly used hypothesis for neural coding. As time goes, our system gradually changes from a quasi-uniform quantizer to a highly non-linear one. Besides, high magnitude stimuli are well refined, while small magnitudes are coarsely approximated. This yields an original bioinspired quantization system, the behavior of which evolves dynamically during the time interval of stimuli observation. Here, we present the retina model adapted to a temporal signal. Then, we explore the input/output map of the system and its ability to recover the original signal. Besides, we raise the analogy between this bioinspired system and already well known compandor/quantizer systems used for analog-to-digital converters. Finally, we compare the performances of our quantizer to the dead zone scalar quantizer used in JPEG2000, and show better performances for low rate transmissions.**

I. Introduction

The retina transforms a continuous input current $I(t)$ into a series of impulses termed as *spikes*. The spikes are the elements composing the neural code of the retina and thus, they convey all the information about the visual stimulus to high order cortical areas. Interestingly, all spikes have the same shape and amplitude, yielding a binary-like code. As a consequence, the retina can be considered as a data quantizer. Several hypothesis are discussed in the literature, on how could this spike-based code be deciphered [1]–[3]. Here we consider the so-called *firing rate coding* proposal, which is the most commonly used one. The rate coding assumes that, given a ΔT-sized time bin, the count of spikes convey the major part of the stimulus information [4].

In order to experiment the behavior of the retina as a quantizer, we implement a three-staged system based on a biologically realistic retinal model of introduced in [5]. The considered simulator is one of the most complete ones generating a spike-based output which, furthermore, successfully reproduce actual neurophysiologic recordings. The model maps the anatomical structure of the retina. This structure is strongly related to the retina functional architecture. Indeed, the retina is a succession of layers. The output of each one is the input of the following. The progression of light stimuli, from the outermost light receptors layer, to innermost ganglionic layer, involves several processing mechanisms. Thus, there are five distinct types of neurons, each one tiling the whole surface of the retina, along a given layer. This architecture is outlined in Figure 1.

As discussed earlier, the innermost ganglion cells layer of the retina emit spikes to convey information over the optic nerve [1].

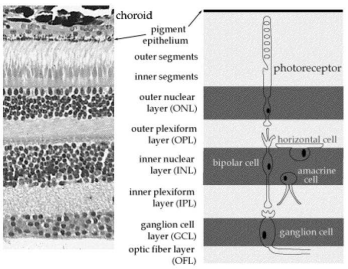


Fig. 1. Diagram showing the layers of the retina. At the top, the outer most pigmented layer. At the bottom, the innermost ganglionic layer (from [6]).

By opposition, retinal cells of the outer stages do note fire spikes. As the input stimuli gets through these stages, the input signal is filtered, but still has the form of a graded continuous electrical signal. These cells are, from innermost to outermost, amacrine cells (in the inner plexiform layer), the Bipolar cells (in the outer plexiform layer), the horizontal cells and finally light receptors (see Figure 1). Only the ganglion cells are responsible for signal discretization. In the following we focus on the cells of the three deepest retina layers, involved in the generation of the visual neural code, namely bipolar, amacrine and ganglioon cells. These cells form the main stages responsible for the shaping of the spiking retina code.

The paper is organized as follows: In Section II, we present the model of the three-staged system. Then, in Section III,we explore the input/output relation of the system and its ability to recover the original signal and, besides, we raise the analogy between our bioinspired coding/decoding schema and well established companding/quantization approaches in telecommunication systems. Finally, in Section IV, we show the performances of the proposed model in terms of rate/distortion trade-off.

II. Biologicaly realistic retina model

Our aim is to study the input/output transfer function of the deep vertebrate retina layers. For this to be done, we base our work on the biologically realistic retina model introduced in [5]. We restrain our study to the temporal behavior of the retina, thus the spatial filtering blocks are ignored in the following description. Furthermore, only the three deepest retina layers in the model are considered, as they are the main stages responsible for the shaping of the spiking retina code. The three-staged simplified model, as implemented in this work, is described through Sections II-A to II-C.

A. Bipolar cells layer: The gain control stage

An issue encountered by any biological system is to adjust its operational range to match the input stimuli magnitude range [2]. Interestingly, fast magnitude adaptation mechanisms are largely observed in the bipolar cells. The input of the bipolar cells stage is a current $I(t)$, and its output is a potential denoted $V_B(t)$.

We suppose, in the following, that $I(t)$ is a time-binned stimulus, such that each sample $I_j = I(t = t_j)$ is exposed to the system during a ΔT-sized time interval. ΔT resolution is sufficiently refined to consider that stimulus is constant during $[0, \Delta T[$. We, then, explore the time course behavior of each sample stimulus signal, defined by:

$$I(t) = \begin{cases} I_j, \text{ if } t \in [0, \Delta T[\\ 0 \text{ otherwise} \end{cases} \quad (1)$$

The gain control procedure, as introduced in [5], is defined by:

$$\frac{dV_B(t)}{dt} + g_B(t)V_B(t) = I(t), \quad (2)$$

where g_B represents a variable leakage term. The expression of g_B, for a potential $V_B(x, y, t)$ that varies in time and space, is given by:

$$g_B(x, y, t) = G_{\sigma_B}(x, y) * E_{\tau_B}(t) * Q\left(V_B(x, y, t)\right), \quad (3)$$

where G_{σ_B} is a spatial filter, E_{τ_B} is a low-pass temporal filter, and Q a static function of V_B. As the input $I(t)$ is a function of time only, assumption is made that the spatial filter $G_{\sigma_B}(x, y)$ is a Dirac impulse, i.e, $G_{\sigma_B}(x, y) = \delta_{0,0}$. E_{τ_B} is a temporal undershoot modeled using exponential filters, and defined by:

$$E_{\tau_B}(t) = \frac{1}{\tau_B} exp^{\frac{-t}{\tau_B}}. \quad (4)$$

Furthermore, Q is assumed to have a convex shape, and is defined by:

$$Q(V_B(t)) = g_B + \lambda_B V_B^2(t). \quad (5)$$

This yields the general developed expression relating the output, V_B, to the input, I, that follows:

$$\begin{aligned} &\frac{dV_B(t)}{dt} - g_B\left(e^{-\frac{t}{\tau_B}} - 1\right)V_B(t) + \\ &\frac{\lambda_B}{\tau_B}\left(\int_0^t V_B^2(t-s)e^{\frac{-s}{\tau_B}}ds\right)V_B(t) \quad = \quad I(t). \end{aligned} \quad (6)$$

As the retina has a layered architecture, the output voltage V_B, of the bipolar cells, is the input of the subsequent amacrine cells in the inner plexiform layer stage (IPL).

B. Inner plexiform layer: The non-linear rectification stage

We consider the signal, of voltage V_B, as generated by bipolar cells of the retina. This current is subject to a non-linear rectification by amacrine cells in the IPL. The output of the IPL is a corrected current I_G. A biologically realistic model of this rectification [5] is defined by:

$$I_G(x, y, t) = G_\sigma(x, y) * N\left(\varepsilon\, T_{w_A,\tau_A}(t) * V_B(x, y, t)\right). \quad (7)$$

In (7), ε, w_A, and τ_A are constant scalar values, T_{w_A,τ_A} is a linear transient filtering defined by:

$$T_{w_A,\tau_A} = \delta_0(t) - \frac{w_A}{\tau_A}e^{-\frac{t}{\tau_A}},$$

and N is a non linear function of V_B defined by:

$$N(v) = \begin{cases} \dfrac{I_A^2}{I_A - \lambda_A(v - V_A)}, \text{ if } v < V_A \\ I_A + \lambda_A(v - V_A), \text{ if } v \geqslant V_A, \end{cases}$$

where I_A, V_A, and λ_A are constant scalar parameters. As only the time behavior is taken into account, the spatial filter $G_\sigma(x, y)$ is set to a Dirac impulse (cf. Section II-A). Developing the expression in (7), we get the following expression for I_G:

$$I_G(t) \quad = \quad N\left(\varepsilon\left(V_B(t) - \frac{w_A}{\tau_A}\int_0^t V_B(t-s)e^{\frac{-s}{\tau_A}}ds\right)\right).$$

I_G is the current input of the last retina stage, namely the ganglionic layer, which originates the neural code of the retina.

C. Ganglion cells layer: The spike generation stage

The ganglionic layer is the deepest one tiling the retina. The ganglion cells are the neurons that generate the spiking output of the retina. A formalization for spike generator neurons in the retina is proposed in [5]. The model chosen is the widely used leaky integrate and fire (LIF) [3]. $I_G(t)$ is the input stimulus of this spike generator layer, and $V(t)$ is its output voltage. We study $V_G(T)$ behavior in the time bin $[0, \Delta T[$, which amounts studying the spike emission timings $(T_i)_{i\geqslant 0}$. $(T_i)_{i\geqslant 0}$ are defined by the following:

$$\begin{cases} V(T_i) \quad = \quad \delta, \forall\, i \geqslant 0,\ T_i \in [0, \Delta T[, \\ V(t) \quad = \quad 0, \forall\, i \geqslant 0,\ \forall\, t \in [T_i, T_i + T_{ref}[. \end{cases} \quad (8)$$

where δ is the integration threshold of the neuron, and T_{ref} its refractory time. In the following, the refractory time will be neglected as $T_{ref} \ll \Delta T$.

Whenever the voltage V reaches δ, the neuron fires a spike, then reinitializes its voltage to 0. Once the spiking mechanism is specified (cf. (8)), the model defines the behavior of $V_G(t)$ in the time bin $[T_i, T_{i+1}[$, as $V_G(t)$ obeys to the following differential equation:

$$c_G\frac{dV(t)}{dt} + g_G V(t) = I_G(t), \forall\, t \in [T_i, T_{i+1}[, \quad (9)$$

where g_G is a constant conductance, and c_G is a constant capacitance. Then, solving (9), we get:

$$V_G(t) = \left(\frac{1}{c_G}\int_{T_i}^{T_i+t} I_G(s)e^{\frac{g_G(s-T_i)}{c_G}}ds + C\right)e^{-\frac{g_G(t-T_i)}{c_G}}, \quad (10)$$

where C is an integration constant. In this section, we specified the model transform that leads to the generation of spikes, here restricted to the time transform. In Section III, we introduce a bio-plausible coding scheme, and specify the corresponding decoding procedure.

III. Neural quantizer based on rate coding

In Section II, we presented a biologically realistic three-staged model for spike generation in the retina. Our aim, in this Section, is to get the overall model characteristic input/output behavior. In order to do this, we first consider each stage transfer function separately, then we cascade the three of them, and finally propose a possible decoding algorithm to recover initial input.

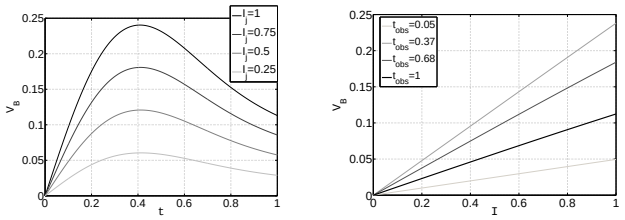
A. Coding pathway

An interesting feature we emphasize in this model, is its dynamics as it involves time t. Our approach is to study each stage, for a given observation time $t = t_{obs}$, then explore how this behavior evolves as t_{obs} varies.

1) The gain control stage: First stage in the model maps each stimulus sample $I(t = t_j)$ (cf. (1)) into a voltage V_B (cf. (6)). For a given maximum stimulus value I_j (cf. (1)), we solve the differential equation (6). The algorithm we use is the Runge-Kutta ordinary differential equation solver [7]. The resulting solution $V_B(t, I_j)$ is shown in the Figure 2(a), for different values of I_j.

For a given observation time t_{obs}, we infer a map that associates a bipolar potential $V_B(t_{obs}, I)$, to an input current I. In order to do this, we observe the system response to a stimulus I at $t_{obs} \leqslant \Delta T$, then explore $V_B(t_{obs}, I)$ for all possible $I = I_j$ values. This leads to the mappings shown in Figure 2(b).

These results prove that, under the assumption of a slot input I (cf. (1)), the gain control in the bipolar cells layer is linear. The linear slope $G_{t_{obs}}$ of the gain is dependent on the observation time t_{obs} of the stimulus.



(a) (b)

Fig. 2. 2(a): $V_B(t)$: Output potential of bipolar cells obtained with Runge-Kutta numerical resolution of (6). Solutions are computed for different input values from $I_j = 1$ (thick line) to $I_j = 0.25$ (thin line). Simulations are made for the following parameter values: $\Delta T = 1$, $g_B = 10^4$, $\tau_B = 10^3$, $\lambda_B = 10^1$.. 2(b):$V_B(I)$: A one-to-one map associating each input current I to a bipolar output potential V_B. Maps are shown for different observation durations t_{obs}, ranging from $t_{obs} = \Delta T$ (thick line) to $t_{obs} = \frac{\Delta T}{20}$ (thin line).

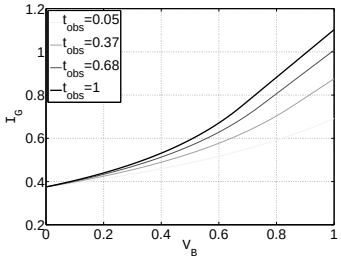


Fig. 3. $I_G(V_B)$: Non linear IPL rectification mapping each V_B value into an output current I_G. Maps are shown for different exposition durations t_{obs}, ranging from $t_{obs} = \Delta T$ (thick line) to $t_{obs} = \frac{\Delta T}{20}$ (thin line).

2) The non-linear rectification stage: After the stimulus is rescaled in the gain control stage, it gets non-linearly rectified in the second IPL stage. Computing the transform in (7), we obtain the mappings shown in the Figure 3, each one corresponding to an observation time t_{obs}. It appears that for short observation times, input is quasi-linearly rescaled, while for longer observation times non linearity is accentuated. This implies that, the instantaneous behavior of the IPL stage is a linear gain control, while as observation goes on, emphasize is made on the high amplitude IPL inputs.

3) The spike generation stage: the rate coding approach: The current I_G, that is generated by the IPL stage, passes through the ganglionic stage yielding a spike-based code. Here we consider the so-called *rate coding* hypothesis, to interpret the coding mechanism of the retina. This is the most commonly used theory. The rate coding assumes that, in a given predefined time bin ΔT, the count of spikes convey the major part of the stimulus information [4].

Through the two preceding stages, input current I is rescaled by a static gain control slope $G_{t_{obs}}$ and corrected by a static non linear function. Thus I_G is supposed constant over the time interval $[0, t_{obs}] \subset [0, \Delta T[$. This assumption is bio-plausible for a sufficiently restrained observation time t_{obs}. We further assume that initial conditioning of V in $[T_i, T_{i+1}[\subset [0, \Delta T[$ (cf. (8)) implies $C = 0$ (cf. (10)). Then for a given $t \in [T_i, T_{i+1}[$, we get:

$$V_G(t) = \left(\frac{I_G}{c_G} \int_{T_i}^{T_i+t} e^{\frac{g_G(s-T_i)}{c_G}} ds \right) e^{-\frac{g_G(t-T_i)}{c_G}}, \forall\, t \in [T_i, T_{i+1}[$$

implying:

$$V_G(t) = \frac{I_G}{g_G} \left(1 - e^{-\frac{g_G(t-T_i)}{c_G}} \right). \tag{11}$$

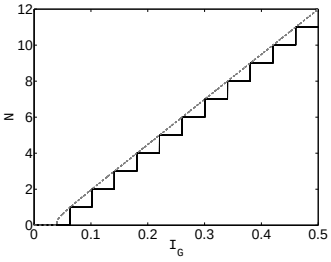


Fig. 4. The rate code generated by the ganglion cells: the rate code $N = f(I_G)$ (solid line) and the real spiking frequency $F = f(I_G)$ (dotted line). Simulations are made for the following parameter values: $\delta = 4$, $g_G = 10^{-2}$, and $c_G = 10^{-2}$.

Under the assumptions we made, the behavior of the ganglion cell, between two successive spikes is the same. Thus $V_G(t)$ is a periodic function of time, and finding the firing timing T_{i+1}, knowing T_i, is equivalent to the deduction of the period of $V_G(t)$. This yields the following formula for the period computation:

$$(T_{i+1} - T_i) \quad = \quad \frac{c_G}{g_G} log \left(\frac{I^c}{I^c - g^L \delta} \right). \tag{12}$$

For further details about proofs, interested readers may refer to [3]. Through (12), we deduce the spiking frequency is $F = \frac{1}{T}$. The rate code, in $[0, \Delta T]$, is the number of spikes fired is N, defined by:

$$\begin{aligned} N &= \lfloor F \Delta T \rfloor \\ N &= \left\lfloor \frac{g_G \Delta T}{c_G \, log \left(1 + \frac{g_G \delta}{I_G - g_G \delta} \right)} \right\rfloor . \end{aligned} \tag{13}$$

We compute the function in (13) for different values of I_G. The result in Figure 4, show the ganglion cell to be a quasi-uniform scalar quantizer after a very short transitory stage around zero.

Based on a biologically realistic model of the retina, we have defined now a rate coding scheme for temporal signals. We propose a possible decoding algorithm in Section III-B.

B. Decoding pathway

Our aim, in this Section, is to recover $\tilde{I}$, the estimation of the input I, knowing its rate code N, and the model parameters. Though the coding scheme in Section III-A is strongly related to actual biological retina behavior, we do not claim that the proposed decoding algorithm is the one that is actually employed in the visual cortex.
The decoding algorithm goes exactly the opposite way of the coding one, from the reverse ganglionic layer to the reverse gain control.

First, we recover $\tilde{I}_G$, the estimation of I_G (cf. (7)). For this, we apply the following reverse mapping:

$$\tilde{I}_G \quad = \quad \frac{g_G \delta}{1 - e^{-\frac{g_G \Delta T}{c_G N}}}. \tag{14}$$

Second, we recover $\tilde{V}_B$, the estimation of V_B, knowing $\tilde{I}_G$. For this, we infer the reverse IPL stage mapping through a look up table. The voltage $\tilde{V}_B$, corresponding to values of $\tilde{I}_G$ that do not match the table elements, are computed by spline interpolation.

Finally, we recover the input signal $\tilde{I}$, by the reverse bipolar gain control. As the gain control in the first coding stage is linear, the reverse gain control is a simple division.

Obviously, the recovered signal $\tilde{I}$ does not match exactly the original I. This is due to the floor operator in the spike generation

mechanism (cf. (8)). The behavior of the coder/decoder system is, thus, analogous to a quantizer/ de-quantizer. We investigate the characteristic behavior of the bioinspired quantizer, we defined, in Section III-C.

C. The retina as a bioinspired quantizer: the overall system behavior

Let us cascade the three layers of our system. We aim at defining the characteristic behavior of the bioinspired quantizer as defined in Sections III-A and III-B, and explore the evolution of it across time. It appears that as the observation time t_{obs} increases, our system goes from coarse to fine, and from uniform to non-uniform.

The refining is intuitive and confirmed by actual neurophysiologic experiments. Indeed the visual cortex perceives global aspects of the stimulus first, then as time goes acquire more information about sharp features.

Then the model quantizer is non-uniform. High magnitude signals are mapped accurately, by a small quantization step, while small magnitude signals are coarsely rendered. This is due to the non-linear rectification in the IPL stage. Indeed, this rectification compresses the dynamic range of small magnitude signals around zero and span higher ones in a linear fashion, this before the generation of spikes in the ganglion cells. This non-uniformity tendency is accentuated as the gain control gets higher across time.

Figure 5 shows an example map of a reconstructed input $\tilde{I}$ as a function of an input I, using the bioinspired quantizer, and this at two different observation timings.

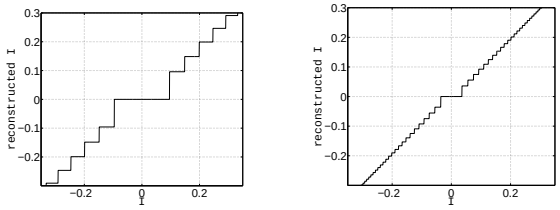


Fig. 5. Input signal/reconstructed signal characteristic: Behavior evolution of the cascaded three stages of the bioinspired model. On the left $t_{obs} = 0.075\,\Delta T$, on the right $t_{obs} = 0.27\,\Delta T$.

Yet, telecommunication systems are already implemented for dynamic signal range compression, namely compandor circuits. Companding is a technique that is widely used in telecommunication [8] making the quantization steps unequal, as the IPL stage does in our case. It is also interesting to denote that companding is preceded, for audio recordings, by a variable-gain amplifier, which is locally linear, in the same manner as the bipolar cells gain control loop described above.

IV. Experimental results

Our goal, in this Section is to test the performances of the bioinspired system in terms of rate/distortion trade-off. In order to do so, we chop the time axis into δt-sized time bins centered around different observation timings τ_j. For each τ_j, the bioinspired quantizer has a different behavior, and thus yields a different neural code $(N_i)_{i>0}^{\tau_j}$. Here, we made the assumption of a firing rate code. $(N_i)_{i>0}^{\tau_j}$ is, thus, a series of *spike counts* integers. We, then, estimate the rate of $(N_i)_{i>0}^{\tau_j}$, and the distortion of the reconstruction $\tilde{I}(t)$ that it allows. Our quantizer is tested for an iid Gaussian 1D signal $I(t)$. The rate of the neural code is estimated by its entropy H^{τ_j}, defined by:

$$H^{\tau_j} = -\sum_i p(N_i^{\tau_j}) log(p(N_i^{\tau_j})). \quad (15)$$

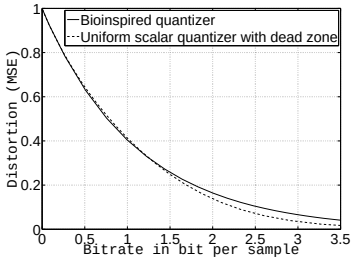


Fig. 6. Rate/distortion trade-off: Comparison between the uniform scalar quantizer with dead zone (dotted line) and the bioinspired quantizer with gain control.

The quality of the reconstruction is measured in terms of mean squared error (MSE) between the original signal I and the reconstructed one $\tilde{I}$. We compare these measures to those of a scalar uniform quantizer with dead zone. Such a quantizer is already a well established standard, used in JPEG2000, and has been proven to be well suited for Laplace-like distributed sources. The results are shown in the Figure 6.

The behavior of the bioinspired quantizer shows slight improvement of the performances for low rate transmission, compared to the scalar uniform quantizer with dead zone. Although, for median rates the classical scalar uniform quantizer is clearly more efficient.

V. Discussion

We presented a bioinspired quantizer based on rate coding. The system implemented relies on a biologically realistic model of the retina. Interestingly, the behavior of the quantizer we specified is similar to an analog-to-digital converter with a companding stage, i.e., involving a non-linear rectification before applying a uniform quantizer. Though, two major differences are to be mentioned. First, the bioinspired model quantizers emphasizes high magnitude signals, while classical approaches aim at refining with more accuracy low magnitudes, which are more probable (Lloyd-Max quantizer). Second, the time dimension has been introduced in the quantization mechanism. This allows further work to explore dynamics of the rate/distortion trade-off and add time scalability to future bioinspired quantization models. Future studies will take into account spatial synaptic dependencies and go a step further toward biological model realism. The final goal of our investigations is to infer a possible decoding procedure for actual neural recordings.

References

[1] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code*, The MIT Press, Cambridge, MA, USA, 1997.

[2] F. Rieke, "Temporal contrast adaptation in salamander bipolar cells," *Journal of Neuroscience*, vol. 21, no. 23, pp. 9445–9454, December 2001.

[3] W.Gerstner and W.Kistler, *Spiking Neuron Models : Single Neurons, Populations, Plasticity*, Cambridge University Press, 2002.

[4] E.D. Adrian, "The impulses produced by sensory nerve endings," *Journal of physiology*, vol. 61, no. 1, pp. 49–72, March 1926.

[5] A. Wohrer and P. Kornprobst, "A biological retina model and simulator, with contrast gain control," *Journal of Computational Neuroscience*, vol. 26, no. 2, pp. 219–249, April 2009.

[6] D.W. Molavi, J. Price, H. Burton, and D. Van Essen, "The w.u.s.m neuroscience tutorial," Web: `http://thalamus.wustl.edu/course/`.

[7] J.R. Dormand and P.J. Prince, "A family of embedded runge-kutta formulae," *Journal of Computational and Applied Mathematics*, pp. 19–26, 1980.

[8] A.B. Clark et al, "Electrical picture-transmitting system," US patent assigned to AT& T, 1928.